

Control for Knowledge-based Information Retrieval

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH
for the degree of
Doctor of Technical Sciences

presented by
Alfred Georg Helmut Ultsch
Diplom-Informatiker TU München
born November 2, 1954
citizen of the Federal Republic of Germany

Accepted on the recommendation of
Prof. Dr. H.-J. Appelrath, examiner
Prof. Dr. C.A. Zehnder, co-examiner

Zürich 1987

Abstract

Information retrieval can be defined as the extraction of specific information out of a great number of stored information items. Information retrieval systems, used for the retrieval of documents, try to answer more or less precise questions about interesting topics with a set of suitable documents or references to documents. Such systems should contain 'knowledge' about the meaning of questions, about the content of the stored information and the particular user's needs for information.

Knowledge-based systems claim to be able to store knowledge and draw conclusions from it. The goal of this thesis is to investigate the use of knowledge-based methods and technologies for information retrieval. A knowledge-based information retrieval system should represent its Information Structures, as well as knowledge in a common knowledge representation formalism. The retrieval process of the system should employ the inferential methods of the used knowledge representation formalism.

A subset of first order logic is chosen for this thesis to represent knowledge. Specially designed retrieval rules represent knowledge for the purpose of retrieval. Retrieval rules capture knowledge about the user's vocabulary, his working domain and his way to perform the retrieval of documents.

The problem of recall and precision of the answers of an information retrieval system is approached by an explicit representation of control knowledge i.e. the knowledge of when and how an interpreter should apply its inferential knowledge. The proposed formulation of control knowledge, in the form of declarative control rules, distinguishes different aspects of the control of the inference process: selection, exclusion, preference and termination.

The representation of control knowledge is one of the basic problems of knowledge-based systems. The main idea of our approach is to retain a declarative representation formalism when control knowledge is concerned and to distinguish different aspects of control knowledge. Some of the aspects of control knowledge can be efficiently implemented (procedurally), while the knowledge representation is declarative, high-level and modular.

The theoretical model of a knowledge-based information retrieval system developed in this thesis specifies the requirements and properties of such a system. In particular, a novel term-similarity function could be defined. Properties like completeness and termination could be derived and boundaries for the amount of overhead of false control strategies could be investigated.

The proposed model is implemented in a prototype of a knowledge-based information retrieval system, called KIR. KIR is a single-user system for personal document and knowledge retrieval running on computer workstations. It is implemented using Prolog and Modula-2.

Zusammenfassung

Information Retrieval kann als das Auffinden spezifischer Informationen aus einer grossen Menge von Informationen definiert werden. Information Retrieval Systeme, welche für das Suchen von Dokumenten benutzt werden, versuchen mehr oder weniger präzise Fragen nach gewissen Inhalten durch Angabe einer Menge zutreffender Dokumente oder Literaturreferenzen zu beantworten. Solche Systeme benötigen "Wissen" über die Bedeutung von Fragen, über den Inhalt der gespeicherten Information sowie über das Informationsbedürfnis des Fragestellers.

Wissensbasierte Systeme erheben den Anspruch, Wissen speichern und aus gespeichertem Wissen Schlüsse ziehen zu können. Ziel dieser Arbeit ist es, wissensbasierte Methoden und Techniken zum Information Retrieval zu untersuchen. Ein wissensbasiertes Information Retrieval System sollte Informationsstrukturen und sonstiges Wissen in einem einheitlichen Formalismus präsentieren. Der Prozess des Auffindens erfolgt dabei mittels Schlussfolgerungsmethoden, die dem Wissensrepräsentationsformalismus inhärent sind.

Zur Repräsentation von Wissen wird in dieser Arbeit eine Teilmenge der Prädikatenlogik erster Stufe gewählt. Wissen über den Retrievalprozess wird dabei in speziellen Retrievalregeln gefasst. Retrievalregeln können u.a. Wissen über den Wortschatz eines Benutzers, über sein Arbeitsgebiet und seine spezielle Art des Suchens von Dokumenten enthalten.

Das Problem von Ausbeute und Präzision der gefundenen Antworten eines Information Retrieval Systems wird durch eine explizite Darstellung von Steuerungswissen (control knowledge), d.h. Wissen darüber, wann und wie inferentielles Wissen anzuwenden ist, angegangen. Die vorgeschlagene Formulierung von Steuerungswissen - in Form deklarativer Kontrollregeln - erlaubt eine Unterscheidung verschiedener Aspekte des Steuerungsprozesses: Selektion, Exklusion, Präferenz und Terminierung.

Die Darstellung von Steuerungswissen ist ein generelles Problem wissensbasierter Systeme. Die wesentliche Idee unseres Ansatzes ist die Beibehaltung einer deklarativen Darstellung von Wissen auf der Ebene der Steuerung sowie eine Differenzierung unterschiedlicher Aspekte von Steuerungswissen. Einige dieser

Aspekte können effizient (prozedural) realisiert werden, trotz einer insgesamt deklarativen, modularen und benutzernahen Darstellungsform.

Das in dieser Arbeit entwickelte theoretische Modell eines wissensbasierten Information Retrieval Systems erlaubt es, Anforderungen und Eigenschaften eines solchen Systems zu beschreiben. Speziell konnte eine neuartige Ähnlichkeitsfunktion für Worte zur Schreibfehlerkorrektur entwickelt werden. Eigenschaften wie Vollständigkeit und Terminierung eines wissensbasierten Information Retrieval Systems konnten präzise gefasst und Schranken für die Kosten des Einsatzes falscher Kontrollstrategien untersucht werden.

Das vorgeschlagene Modell wurde in einem Prototyp eines wissensbasierten Information Retrieval Systems namens KIR realisiert. KIR ist ein Einbenutzer-System für persönliches Literatur- und Wissensretrieval auf Arbeitsplatzrechnern. Es ist in den Sprachen Prolog und Modula-2 implementiert.